A STUDY ON HIGH-PERFORMACE PORTRAIT SEGMENTATION USING THE ENSEMBLE OF DEEP-LEARNING MODELS AND HYBRID MODEL: ACCURACY AND SPEED PERSPECTIVE

Synopsis submitted to CHRIST (Deemed to be University)
as partial requirement for the award of the Degree of

DOCTOR OF PHILOSOPHY
IN
COMPUTER SCIENCE AND ENGINEERING

by

YONG-WOON KIM

Under the Supervision of
Addapalli V N Krishna
Professor

CHRIST
(DEEMED TO BE UNIVERSITY)
BENGALURU · INDIA

Signature of the Research Scholar

Signature of the Research Supervisor

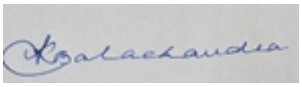
CENTRE FOR RESEARCH
CHRIST(Deemed to be University)
BENGALURU

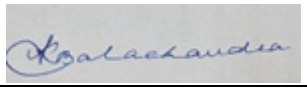
MARCH 2022

CENTRE FOR RESEARCH

Approval of PhD Synopsis

[as per clause 11(k.) of CU Regulation for the Degree of Doctor of Philosophy, May 2019]

Name of Scholar:	Yong Woon Kim
Department:	Department of Computer Science and Engineering
Name of Co Supervisor:	
Names of Research Advisory Committee: 1. Dr Addapalli V N Krishna 2. Dr Balachandran K 3. Dr Diana Jeba Jingle	
Name and Designation of External Reviewer if applicable:	
Date of Synopsys Presentation:	
Title of the PhD thesis:	A study on high-performance portrait segmentation using the ensemble of deep-learning models and hybrid model: accuracy and speed perspective.
Approved by: 1. Dr Addapalli V N Krishna Signature: 2. Dr Balachandran K Signature:  3. Dr Diana Jeb Jingle Signature:	

Head of Department 

Date of approval by Doctoral Committee:

Director Centre for Research

ABSTRACT

Portrait segmentation (PS) can be referred as a process of segmenting a person in an image from its background. There are many real applications using the PS technology. For example, camera apps of mobile phone and online services support instant PS for selfie photos, online video conferencing solutions provide virtual background function for entertainment, aesthetic, privacy, and security reasons, robot vision requires the PS for human-robot interaction, and the PS can be used for content- based image/video retrieval in database application. A good number of studies show that the Deep-Learning based segmentation model (DSM) is a reasonable choice for the PS. However, the study on the PS using the ensemble of multiple DSMs is a new domain. Unlike the PS using a single DSM, the ensemble method combines the segmentation results of multiple DSMs to generate an optimal result. This work proposes three ensemble approaches for the PS to achieve a high-accuracy and high-speed segmentation for various applications. Firstly, the PS using the ensemble of DSMs for portrait images. Secondly, the PS using the ensemble of DSMs for portrait videos. Lastly, the PS using the ensemble of DSMs and hybrid method for portrait videos. The effect of image downscaling for PS was also evaluated and analysed. This work evaluated the proposed approaches using publicly available datasets named EG1800 for portrait image segmentation and self-constructed portrait video datasets having a single- person view collected from publicly available websites. The proposed approaches were evaluated by measuring accuracy, variance error, bias error, and efficiency.

CONTENTS

ABSTRACT

1. INTRODUCTION 1

2. LITERATURE REVIEW 3

3. PROBLEM IDENTIFICTION & STATEMENT 6

4. OBJECTIVES 7

5. METHODOLOGIES 8

6. SOFTWARES AND TOOLS USED 12

7. ACCOMPLISHMENT OF THE RESEARCH WORK 13

8. ANALYSIS/SIMULATION/TESTING/VALIDATION 15

9. RESULTS AND DISCUSSIONS 17

10. CONCLUSIONS 23

BIBLIOGRAPHY 24

PUBLICATIONS 26

TABLES

Table 1. Summary of Literature Review 3
Table 2. Motivation, requirements, challenges, and gaps of this work 6
Table 3. Objectives table 7
Table 4. Comparison of methods....11
Table 5. Training and testing dataset of models. 15
Table 6. Experimented DSMs. 15
Table 7. Validation of datasets 16
Table 8. Intersection over Union (IoU) value comparison. (%) 17
Table 9. Average of all models (%)....18
Table 10. Average of MER and CER 19
Table 11. IoU values of Parallel-BMDS 20
Table 12. IoU values of Ensemble-BMDS 20
Table 13. Comparison of execution time (millisecond) 22

FIGURES

Figure 1. Ensemble of n-DSMs. 8
Figure 2. Converting binary value to probability value....9
Figure 3. Conceptual diagram for PVS using NF ensemble approach....10
Figure 4. Conceptual diagram of PVS using Parallel-BMDS 11
Figure 5. IoU comparison of single models and ensemble models. 17
Figure 6. Segmented results of single models and Two-Models ensemble: (a) original image; (b) ground truth; (c) single models; and (d) proposed ensemble model. 18
Figure 7. Comparison of IoU and FLOPs; M3M2S@1 = MNV3+MNV2+SN (1F); M2S@1 = MNV2+SN (1F); M2S@2 = MNV2+SN (2F); M3M2S@3 = MNV3+MNV2+SN (3F); M2PS@3 = MNV2+PN+SN (3F); M3S@2 = MNV3+SN(2F). 19
Figure 8. Average of CER and MER....19
Figure 9. (a) Input frames; (b) output of Segmentation mode 1; (c) output of Segmentation model 2; (d) output of proposed N-Frames Ensemble; (e) expected output (ground truth)....20
Figure 10. Speed comparison 21
Figure 11. Segmentation examples of continuous video frames (5 frames); (a) video frames; (b) ground truth; (c) segmentation using a slow DSM; (d) segmentation using Parallel-BMDS 21

1. INTRODUCTION

Portrait segmentation (PS) in general can be referred as a process of segmenting a person in an image from its background. There are many real applications using a PS technology. B612 is one of the most popular camera apps with over 500 million downloads in the Google play store. It can change the background of a photo automatically with one-click. Some online services also provide automatic background change feature. PS is not only for entertainment, but also used for online video conferencing. The Zoom meeting is one of the most popular online video conferencing solutions and the use of it was increased after the COVID-19 pandemic. Many organizations are using Zoom meeting for the virtual conference and online conferencing solutions as many people are working from home. PS technology is used in the Zoom meeting to hide the user's background for privacy, security, and aesthetic reasons. Robot Vision or Machine Vision requires PS technology for the Human-Robot interaction. A robot or machine can segment a human body from a camera image and can process a human information for Human-Robot interaction. For Robot Vision or Machine Vision, the image segmentation should be executed on an embedded hardware. In addition to these, the PS technology can be used for content-based image retrieval. When a user enters a query to find all photos with similar people, the human body of the entered photo can be segmented and passed to the database to find similar people in the database. Recently, Google announced a new feature called "Google photos cinematic 3D" in their Google photos app. It also uses the concept of PS technology.

PS is a subset of image segmentation. Several prominent image segmentation methods have been studied and developed by researchers. Many of them use traditional image segmentation methods, such as thresholding, watershed, region growing, clustering using contour and edge, graph cut, and Markov random fields.[1]-[2]. A segmented area should be homogeneous and uniform for good image segmentation, but it remains a challenging task [3]. Deep-Learning (DL) based Segmentation Model (DSM) has opened a new era of image segmentation. DSMs have made remarkable improvements in speed and accuracy for image segmentation compared to traditional methods [4]-[9]. These models predict segmentation regions using semantic labels for every pixel of the image [10]. In general, an ensemble of multiple methods can improve the performance of image segmentation. The key of the ensemble approach is to combine various models to create a more effective model [11]. The collective decision produced by the ensemble can reduce generalization errors of prediction and improve the performance accuracy [12], [13]. Many studies show that the combination of multiple segmentations can produce better predictive performance than the individual segmentation model [14]-[16]. Video Instance Segmentation (VIS) is a task that can classify objects into the pre-defined classes, trace objects throughout a video, segment classified objects, and classify objects with localization within a video. The extension of instance image segmentation to the video domain, was necessitated due to the increased use of video applications such as online video conferencing solution. Portrait Video Segmentation (PVS) is a subset of VIS. It can recognize and extract the human body, normally the upper body, from continuous video frames [17]-[18].

A. MOTIVATION

There are many applications using PS technology and the use of PS is increasing these days. However, high performance and real-time PS remains a challenging problem even with recent developments on image segmentation and matting using DSMs [5], [18]-[23]. There are several studies on PS using single DSMs for single images. In general, these studies focus on optimization or customization of a DSM architecture. Several studies on the ensemble method of multiple DSMs show better image segmentation accuracy compared to single DSMs. However, single DSMs are not nicely fit for PVS considering the characteristics of a video, such as sudden noises or the continuity of adjacent video frames, and the ensemble approach is not satisfactory when we directly apply it to PVS because the speed of segmentation is slower than a single DSM. PS using the ensemble of multiple DSMs are a new domain to be studied further.

B. CONTRIBUTIONS

To address these issues, this work proposes three novel approaches for PS using the ensemble of multiple DSMs to achieve a high-performance automatic segmentation. Firstly, the PS using the ensemble of heterogeneous DSMs for portrait images. Secondly, the PVS using the N-Frames ensemble of DSMs for portrait videos. Lastly, the PVS using the ensemble of DSMs and hybrid method for portrait videos. The proposed methods are not limited to PS but can also be applied to segment any objects in a static image or a video. The followings are contributions of this study.

[Objective 1]

1) A novel approach of PS using the ensemble of multiple heterogeneous Deep-learning base Segmentation Model (DSM) was proposed.
2) A simple and efficient way to combine the segmentation results of single DSMs was introduced. It is easy to implement and effective to reduce falsely predicted region.
3) Intersection over Union (IoU), IoU standard deviation, False Negative Rate (FNR), False Discovery Rate (FDR), FNR + FDR and |FNR-FDR| were used to measure the accuracy, variance and bias error of segmentation results. IoU standard deviation is used to measure the variance error of experimented models, FNR + FDR to measure the bias error, and |FNR-FDR| to measure the balance of bias error. These metrics are a new approach to evaluate the performance of DSMs and proposed approaches.
4) The efficiency rate of memory and computing power was measured to evaluate the efficiency of DSMs and proposed models. This work attempts to measure cost efficiency of state-of-the-art DSMs, and proposed approaches.
5) The optimal combinations of DSMs were explored and analysed.

[Objective 2]

1) A novel approach of Portrait Video Segmentation (PVS) using N-Frames ensemble of multiple DSMs was proposed.
2) A simple and efficient way to combine the segmentation results of multiple video frames was introduced. It is easy to implement and effective to reduce the sudden noise of video frames.

3) A novel ensemble approach using the rotation method of DSMs for PVS was introduced and experimented.
4) The cost efficiency of single DSMs, N-Models ensemble, and N-Frames ensemble for PVS was measured and compared. The result was categorized into four groups, i.e., high-cost & high-performance, low-cost & high-performance, low-cost & low-performance, and high-cost & low-performance groups.
5) This work constructed a new video dataset to evaluate the performance of PSV.

[Objective 3]

1) A novel approach of PVS using the ensemble of multiple DSMs and hybrid method was proposed.
2) A simple and efficient way to combine the segmentation results of multiple video frames using hybrid segmentation method was introduced.
3) A serial and parallel mode of hybrid segmentation was proposed. It enables to save the energy of inference processing and enables high-speed segmentation even on the low performance hardware.
4) A boundary and edge adjustment algorithms were introduced to resolve the inconsistency of continuous frames.

[Additional]

1) Image processing and portrait segmentation using downscaled images were evaluated and its effect was analysed.
2) Datasets used for PS and edge detection algorithms were explored and analysed.

2. LITERATURE REVIEW

Table 1 shows the summary of literature review. This work collected more than 120 related works and checked more than 1,900 references. After the intensive literature survey, 16 research papers related to deep-learning based PS since 2016 have been found. Out of 16 papers, 5 are related to general object segmentation using the ensemble of DSMs. However, researches for PS using ensemble of multiple DSMs have not been reported.

Table 1. Summary of Literature Review

Related Works	Occurrence
A. More than 120 related works and more than 1,900 references were reviewed.	
B. Studies on Portrait Segmentation (PS) using a single Deep-Learning based semantic Segmentation Model (DSM).	2016 ~ (14 papers until 2021)
C. Studies on object segmentation using an ensemble of DSM.	2016 ~ (5 papers until 2021)
D. Studies on portrait video segmentation using a hybrid method of a DSM and image processing algorithms.	2019 (1 paper)
E. Studies on PS using the ensemble of DSMs.	Not reported

A. IMAGE SEGMENTATION USING DEEP-LEARNING

Long et al. [24] proposed the Fully Convolutional Network (FCN) which uses the existing classification networks such as AlexNet, VGG net, and Google Net into fully convolutional networks, and transfer their learned representations to the segmentation task by fine-tuning. FCN achieves state of-the-art segmentation of PASCAL VOC (20% improvement over 2012), NYUDv2, and SIFT Flow, while inference for a typical image takes less than one fifth of a second. Chen et al. [25] proposed a network architecture called DeepLabv3. It adopts an atrous convolution for the downsampling layer and upsampled filters for a dense feature map extraction and for long-range context capturing. It also present a modified Xception model that improves the performance with faster computation. Mark Sandler et al. [26] introduced MobileNetV2, an improved version of MobileNets which is an efficient and lightweight segmentation model for mobiles and embedded systems. MobileNetV2 is based on an inverted residual structure that connects the thin bottleneck layers, and it improves the state-of-the-art performance on multiple tasks as well as for a range of different model sizes. Andrew Howard et al. [27] proposed MobileNetV3-Large and MobileNetV3-Small models using hardware-aware Network Architecture Search and NetAdapt algorithm. It was developed to achieve the best semantic segmentation on mobile devices. Due to the performance and efficiency of MobileNetV2 and MobileNetV3, these models are used as a backbone with other networks such as DeepLabv3, U-Net, LR-ASPP, etc., for semantic segmentation. Song-Hai Zhang et al. [28] proposed a real-time PS model called PortraitNet for mobile devices. The model includes two modules, the encoder module and the decoder module. PortraitNet utilized MobileNetV2 as a backbone in the encoder module and U-shape architecture as a decoder. It shows remarkable performance in accuracy and efficiency, especially when it comes to generating results with sharper boundaries and under harsh illumination conditions. PortraitNet processes 224 × 224 RGB images at 30 Frame Per Second (FPS) on the iPhone 7. Hyojin Park et al. [29] introduced an extremely lightweight PS model, SINet, which used an information blocking decoder to measure the confidence score. It blocks the flow of the decoder utilizing the information. SINet performs well on mobile devices with 100.6 FPS and 95.29% accuracy. Xiaoyong Shen et al. [30] improved the performance by using new PS dataset for fine-tuning the model. In general, CNNs use the "personal class" for character area extraction using FCN-8s, but it is not sufficient in terms of accuracy. The authors simplified the FCN dataset having the foreground and the background only, and trained the model using new dataset.

B. IMAGE SEGMENTATION USING ENSEMBLE

Warfield et al. [31] introduced an algorithm to combine multiple segmentations and validate image segmentation performance. They used Simultaneous Truth And Performance Level Estimation (STAPLE) approach. STAPLE can accurately estimate the segmentation and performance parameters from the observations of segmentation. Andrew Holliday et al. [32] applied a compressed model technique for DL ensemble to the problem of semantic segmentation. The work achieved real-time performance and have similar accuracy to ensembles. D. Marmanis et al. [33] applied the ensemble of multiple DL models using FCN for semantic segmentation of very high-resolution images. The work achieved state-of-the-art semantic segmentation performance on the ISPRS semantic labelling dataset.

C. OBJECT SEGMENTATION IN VIDEO

J. Li et al. [34] proposed a novel approach to segment an object in a video using a proposal-driven framework. The proposed method can effectively utilize the common information between video frames to improve segmentation accuracy. The authors adopted the ResNet model for the proposal and the PSPNet model for object segmentation. The work achieved higher accuracy in video segmentation of the VOT 2016 dataset, the DAVIS 2017 dataset, and SegTrack v2 dataset. Yifan Liu et al. [35] developed a real-time video segmentation method that considers accuracy and temporal consistency. The proposed method conducts per-frame inference using compact networks. The authors included the PSPNet18, the MobileNetV2 and a lightweight HRNet to verify that the proposed methods can improve the segmentation accuracy and the temporal consistency without extra computation and post-processing during inference. Based on Cityscapes and Camvid experiments, the compact models outperform previous state-of-the-art semantic video segmentation methods by considering accuracy, temporal consistency, and inference speed. Shan Lin et al. [36] proposed a Multi-frame Feature Aggregation (MFFA) module to improve instrument segmentation. It uses temporal and spatial relationships between frame pixels for feature aggregation. The authors applied the proposed approach to the real-time instrument segmentation using DeepLabV3+ with ResNet50 and MobileNet as backbone feature extractors. The proposed MFFA can reduce the deep encoder to its trimmed version and decrease computation costs. MFFA can also be easily combined with any segmentation model that uses an encoder-decoder architecture. Yifan Wang et al. [37] introduced a PVS method that combines a soft correspondence network (SCN) and a portrait segmentation network (PSN). It Automatically segments the dominant person in each frame of a portrait video sequence, while ensuring the temporal consistency of segmentation results. They have achieved high accuracy temporal-coherent segmentation result using a soft correspondence network. One limitation of this work is that the proposed dataset and method are based only on a single target scenario. Tairan Zhang et al. [38] proposed a real-time framework for human segmentation designed specifically for single-person video sequence. The framework combined a CNN model and a visual tracking system using a level set algorithm. The CNN model obtained a human segmentation result from a specific frame in a video and passed it to the tracking system to capture the human segmentation of the rest frames. The work achieved a better balance between segmentation accuracy and speed. One limitation of the approach is that the framework may fail with complex backgrounds.

3. PROBLEM IDENTIFICTION & STATEMENT

Compared to traditional image segmentation methods, these DSMs achieved remarkable performance improvement in image segmentation [5]-[9]. Some studies showed that DSMs could be utilized for a PS purpose. However, these works mainly aim to segment objects from a single image but does not perform well when it is applied to PS in a video. In general, an ensemble model can generate optimized results using the combination of multiple machine-learning models or deep-learning models. Several studies applied the ensemble method using multiple DSMs to the image segmentation domain and achieved better performance compared to single DSMs. These works showed that the ensemble of various DSMs can be a reasonable choice for image segmentation applications. However, these ensemble approaches require more computing power than single DSMs. For the PSV, the segmentation speed is very important. The extension of instance image segmentation to the video domain is a natural step as the requirement of image segmentation in a video is increasing. There are several studies on image segmentation in a video. These studies showed that DSMs could be used for image segmentation in a video. In these studies, deep-learning models for image segmentation were optimized by considering the characteristics of a video. However, the ensemble of multiple DSMs for the PVS is an area that needs further study. Table 2 shows motivation, requirements, challenges, and gaps of this work.

Table 2. Motivation, requirements, challenges, and gaps of this work

Motivation	• Many applications: selfie camera apps, online conferencing solutions, content-based image retrieval, robot-human interaction, etc.
Requirements	• High-accuracy and/or high-speed portrait segmentation for various applications.
Challenges	• High-performance and real-time PS remains a challenging problem even with recent developments on image segmentation using single DSMs. • Several studies on image segmentation using the ensemble of DSMs show better accuracy compared to single DSMs. However, the ensemble approach is not fit well when we directly apply to portrait video segmentation.

Gaps	• A good number of studies show that:
	  ○ DSM is a reasonable choice for the PS (14 related works)
	  ○ Ensemble of multiple DSMs can improve the precision of segmentation (5 related works)
	• However, the PS using the ensemble of multiple DSMs have not been studied.

[Problem statement]

It is evident that DSMs are a reasonable choice for the PS, and the ensemble of multiple DSMs can improve the precision of the segmentation result. But it is not fit well when we apply these approaches directly to the image segmentation in a video. Considering the advantage of the ensemble model, it is an appropriate attempt to utilize the ensemble of multiple DSMs for PVS. The ensemble of multiple DSMs for the PS is an area that needs further study.

4. OBJECTIVES

1) High-performance portrait segmentation for selfie photos using ensemble method
2) High-performance portrait segmentation for portrait videos using ensemble method
3) High-performance portrait segmentation for portrait videos using hybrid method

Table 3 shows three objectives of this work and approach details.

Table 3. Objectives table

	OBJECTIVE-1	**OBJECTIVE -2**	**OBJECTIVE -3**
	PS for selfie photos using ensemble method	PS for portrait videos using ensemble method	PS for portrait videos using hybrid method
HIGH-ACCURACY	*Ensemble of Heterogeneous DSMs*	*Ensemble of N-Frames*	Ensemble of DSMs
HIGH-SPEED	*Optimal combination*	*Rotation method*	Hybrid method
	Downscaling of images	*Downscaling of images*	*Downscaling of images*

5. METHODOLOGIES

This work uses the empirical, quantitative and qualitative methodologies. The following section describes the approaches of this work in detail.

A. Ensemble Method

In machine learning, ensemble models are developed by combining the prediction of multiple individual models to improve the overall predictions. The ensemble model can be created either by combining multiple modelling algorithms or using different training datasets. The main purpose of ensemble method is to reduce the generalization error in the machine learning algorithms. This work proposes the ensemble of *n*-DSMs to enhance the accuracy performance of the single models, where *n* indicates the number of single models combined at a time. Fig. 1 shows the structure of ensemble approach of *n*-DSMs. Consider a set of $\boldsymbol{T}$ number of PS models i.e. $\{\boldsymbol{L}_1, \boldsymbol{L}_2, \boldsymbol{L}_3, ..., \boldsymbol{L}_T\}$, where $\boldsymbol{L}_i$ are individual PS models (First-Level learners). The input images from the EG1800+CDI datasets are fed to the individual models, and the segmented output masks are obtained and stored. These individual segmented output masks are combined to enhance final outputs. For the ensemble of the outputs of individual models, the simple soft voting method and weighted soft voting method are used in this research. The idea of the ensemble is to learn from the outputs of individual models, and collectively produce a better segmented result.

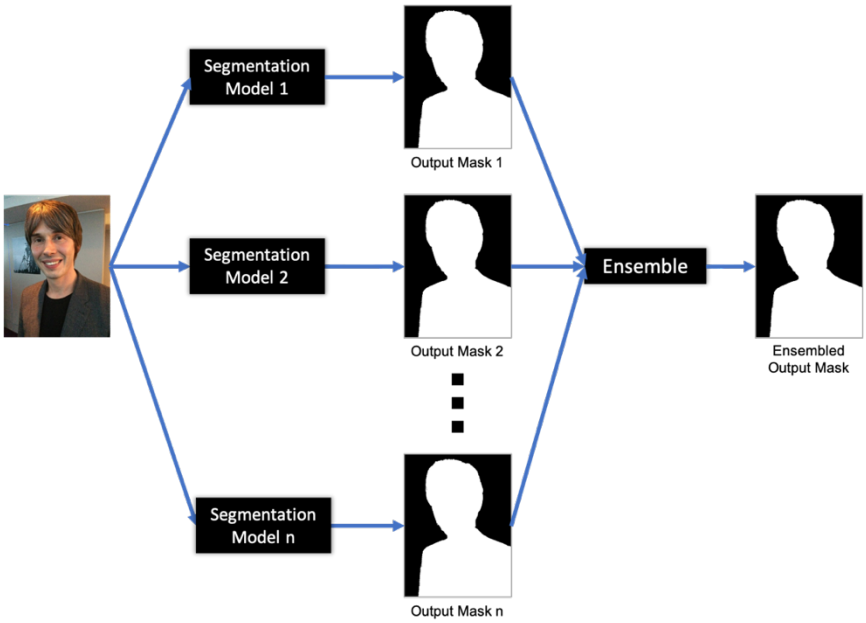


Figure 1. Ensemble of n-DSMs.

In this work, simple soft voting and weighted soft voting as a Meta-Classifier in a stacking ensemble were used to combine the segmented output masks of individual models. In the simple soft voting method using Equation (1), the weights of outputs of individual models are treated equally. In weighted soft voting using Equation (2), the weights of outputs of individual models are not equal, and are trained from the result of First-Level classifiers in the stacking ensemble to find a respective weight ratio.

$$H^j(x) = \frac{1}{T}\sum_{i=1}^{T} h_i^j(x) \quad (1)$$

$$H^j(x) = \sum_{i=1}^{T} w_i h_i^j(x) \quad (2)$$

In this work, the segmented output mask generated from individual models have a binary value, i.e., 1 or 0. This binary value was converted into probability value by applying the Gaussian Blur function. Gaussian Blur is commonly used in image processing to smoothen and reduce the noise in the image. Gaussian Blur uses a non-uniform low pass filter that preserves the low spatial frequency, and reduces the noise and negligible details in the image. It belongs to linear smoothing filters class with weights according to the shape of a Gaussian function [31]. In this paper, Gaussian Blur is used to convert the binary image mask into the probability distribution, which indicates how much the pixel is close to the segmented human region. Fig. 2 shows the effect of applying Gaussian Blur to a binary image mask. The left metrics shows 9 × 9 pixels divided into two values, 1 or 0. If the pixel value is 1, then it belongs to the segmented foreground region. If the pixel value is 0, then it belongs to the background region. The right metrics shows that the binary pixel values are converted into probability values after applying Gaussian Blur using a 5 × 5 kernel. If the pixel value is closer to 1, then it has higher probability to belong to the segmented foreground region. Converting the binary pixel value into the probability value is also useful to calculate Equations (1) and (2).

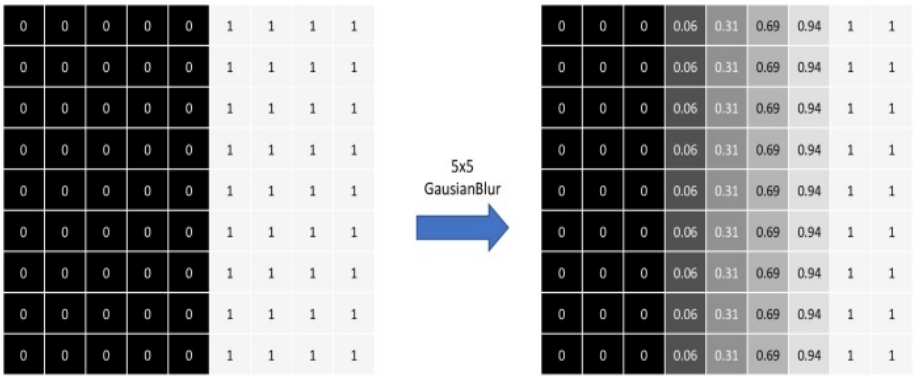


Figure 2. Converting binary value to probability value.

B. N-Frames Ensemble

This paper proposes a novel ensemble approach for PVS called N-Frames (NF) ensemble. Fig. 3 shows the conceptual diagram for PVS using the NF ensemble approach. The NF ensemble method uses a single segmentation model to generate an output result for a single frame in a video. It combines the output results of previous frames with the current output result using the ensemble method. In detail, at a given time t, a video has a frame t, and the frame t is fed to a segmentation model n to generate a segmented n^{th} output mask. The NF ensemble method combines the output mask with other output masks generated from previous segmentation model n-1, …, model 1 to generate the final mask t. It rotates each segmentation model using a round-robin way. For example, if there are total n models for ensemble and model n was used for the frame t at a given time t. The NF ensemble uses model 1 for the frame t+1 at a given time t+1.

It combines the segmented output mask of model 1 with previous output masks generated from segmentation model n, …, model 2 to create the final output mask t+1.

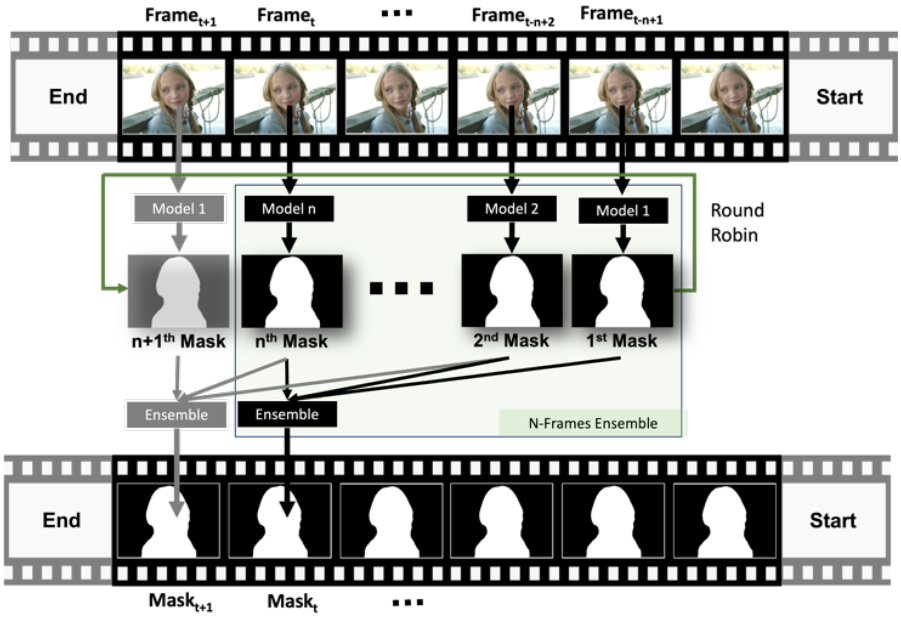


Figure 3. Conceptual diagram for PVS using NF ensemble approach

C. Hybrid method

This work proposes a hybrid method using the combination of a DSM and a Boundary Movement Detecting Segmentation (BMDS) method. The BMDS method has two models called the Serial-BMDS and the Parallel-BMDS. The Serial-BMDS proceed a DSM first followed by BMDSs processing sequentially. It is suitable for fast DSMs which can complete its segmentation inference in a frame time. The Parallel-BMDS proceed a DSM and BMDSs concurrently. It is suitable for slow DSMs which requires longer time than a frame time to complete its segmentation inference. Fig. 4 shows the conceptual diagram of PVS using the combination of DSM and the Parallel-BMDS. Proposed approach uses a DSM to generate a base segmentation mask called a "base-mask" from a current video Frame$_{t+1}$. It passes the base-mask of a current video Frame$_{t+1}$ to the BMDS process for a video Frame$_{t+m}$. The BMDS method adjusts the base-mask of a video Frame$_{t+1}$ using a traditional Background Subtraction (BGS) algorithm and an edge-detection algorithm to produce a segmentation mask of a current video Frame$_{t+m}$. The segmentation mask generated from the BMDS process becomes a new base-mask for the next video Frame$_{t+m+1}$ and the process pass the new base-mask to the next BMDS process. Proposed approach applies BMDS method *n*-times to generate segmentation masks until Frame$_{t+m+n.}$ This work assumes BMDS can complete its process within a frame time. Parallel-BMDS processes current video frames while DSM is generating a segmentation mask. The proposed approach utilises a multi-threading or a multi-processing to handle BMDS simultaneously while DSM is inferring. The advantage of the Parallel-BMDS is an ability to produce segmentation masks of current video frames without missing while DSM is busy for generating a segmentation mask of a previous video frame.

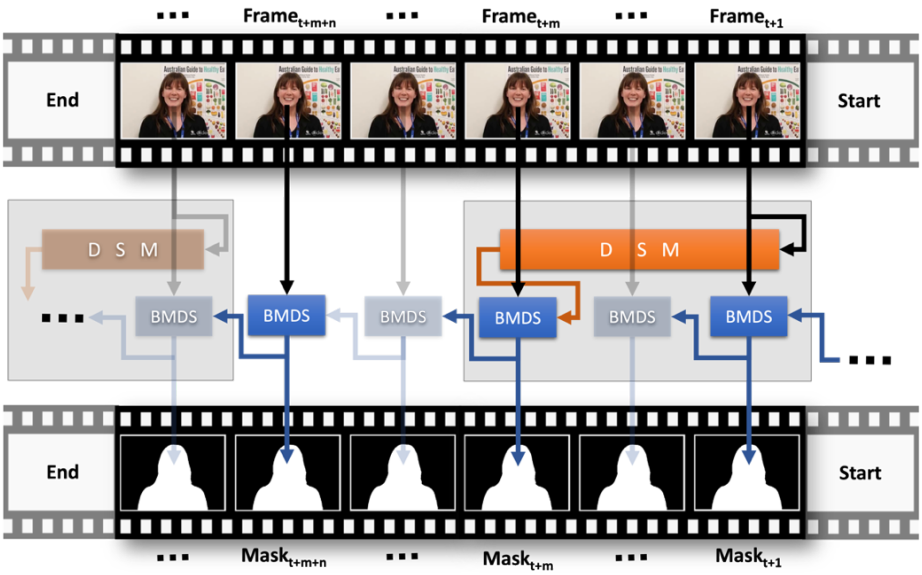


Figure 4. Conceptual diagram of PVS using Parallel-BMDS

BMDS method uses BGS and Canny Edge Detection algorithms. Frame difference is one of the simplest algorithms for BGS. Its equation is as below:

$$P[F(t)] = P[I(t)] - P[B]$$

Mean filter also can be used for BGS. Its equation is as below:

$$B(x, y, t) = \frac{1}{N} \sum_{i=1}^{N} V(x, y, t - i)$$

Table 4 compares the approaches of related works and this work.

Table 4. Comparison of methods

		Related Works	This Work
OBJ-1	**Speed**	• Optimization of DSM architecture (11 papers).	• Optimal combination of DSMs.
	Accuracy	• Optimization of DSM architecture. • Ensemble of DSMs for medical or scene segmentation (5 papers).	• Ensemble of DSMs for portrait image segmentation.
OBJ-2	**Speed**	• Optimization of DSM architecture (3 papers).	• Rotation of multiple DSMs.
	Accuracy	• Optimization of DSM architecture.	• N-Frames ensemble of DSMs.
OBJ-3	**Speed**	• Hybrid method for high-performance hardware (1 paper). • Used level-set method. • No consideration for speed of DSM.	• Hybrid method for low-performance hardware. • Frame difference method. • Serial and parallel mode considering the speed of DSM.
	Accuracy	• Object tracking using base-mask of DSM.	• Boundary adjustment using base-mask of DSM.

D. Measurement Metrics

The Intersection over Union (IoU) is a metric used to measure the accuracy of image segmentation. An IoU is defined as in equation (3) below [28]:

$$IoU = \frac{A \cap B}{A \cup B} \quad (3)$$

where A is an image segmentation result, B is a ground truth image, A ∩ B is the intersection of A and B, and A ∪ B is the union of A and B. IoU standard deviation is used to measure the prediction variance error.

False Negative Rate (FNR) and False Discovery Rate (FDR) are used for validating the accuracy measurement. FNR is used to measure lesser regions than the ground truth. FDR is used to measure larger regions than the ground truth. FNR and FDR are defined as in (4) and (5) below:

$$FNR = \frac{FN}{FN + TP} \quad (4)$$

$$FDR = \frac{FP}{FP + TP} \quad (5)$$

where FN is False Negative, TP is True Positive, and FP is False Positive. The bias error means the amount of difference between the ground truth and the prediction. In general, the ensemble of multiple models can reduce bias error. In this paper, FDR+FNR and |FDR-FNR| is used to measure the bias error of prediction.

In this paper, Memory Efficiency Ratio(MER) and Computing Efficiency Ratio(CER) are used to measure the cost efficiency rate of segmentation models. MER indicates the memory efficiency of the model. Since an efficiency ratio can be denoted as costs over gain, MER measures the required memory size to gain accuracy, and it is calculated as in (6).

$$MER = \frac{M}{IoU} \quad (6)$$

where M is the number of parameters and IoU is the accuracy of a given model.
CER indicates the computing power efficiency of a model. CER measures required computing power to gain accuracy. CER is calculated using (7) as below:

$$CER = \frac{C}{IoU} \quad (7)$$

where C is Floating-Point Operations (FLOPs), and IoU is the accuracy of a given model.

6. SOFTWARES AND TOOLS USED

- Python programming language.
- OpenCV library for image processing.
- TensorFlow, PyTorch, Keras library for deep-learning models.
- Graphic editing tools for portrait image/video editing and annotation.

7. ACCOMPLISHMENT OF THE RESEARCH WORK

This work collected more than 120 related works and checked more than 1,900 references. After the intensive literature survey, 16 research papers related to deep-learning based PS since 2016 have been found. Out of 16 papers, 5 are related to general object segmentation using the ensemble of DSMs. However, researches for PS using ensemble of multiple DSMs have not been reported.

This work measured and compared the performance of segmentation models using Intersection over Union (IoU) for accuracy, IoU standard deviation for variance error, false prediction rate for bias error, and Memory Efficiency Rate and Computing power Efficiency Rate for efficiency measurement.

The following table describes the accomplishment and result of the proposed three approaches.

Objective-1		Approach	Accomplishment & Result
PS for selfie photos using ensemble method	**Speed**	Optimal combination of DSMs.	The optimal combinations of DSMs were explored and experimented. Three models ensemble (MNV3+PN+SN) showed 96.5% accuracy using 2.7 Giga Floating Point Operations (FLOPs) computing power while MNV2 showed 96.2% accuracy using 7.2 Giga FLOPs.
	Accuracy	Ensemble of heterogeneous DSMs	A novel approach using the ensemble of heterogeneous DSMs for portrait image segmentation was proposed and evaluated using the EG1800 dataset. The proposed method was superior to single DSMs in terms of accuracy, bias error and variance error. It showed more than 1.4% higher accuracy than single DSMs on average.

Objective-2		Approach	Accomplishment & Result
PS for portrait videos using ensemble method	**Speed**	Rotation of multiple DSMs.	A novel approach using the rotation method of DSMs for portrait video segmentation was introduced and experimented. The required FLOPs of PS processing was 2.49 Giga FLOPs for both the proposed method and single DSMs on average.
	Accuracy	N-Frames ensemble of DSMs.	A novel N-Frames ensemble of DSMs for portrait video segmentation was proposed and evaluated. The proposed approach performed with more than 2% higher accuracy than single DSMs on average. It also showed lower variance and bias error than single DSMs.

Objective-3		Approach	Accomplishment & Result
PS for portrait videos using hybrid method	**Speed**	Hybrid method	A novel approach using the hybrid method for portrait video segmentation was introduced and evaluated. The proposed approach achieved 5x faster segmentation speed than a single DSM while keeping approximately 99% accuracy rate compared to a single DSM.
	Accuracy	Ensemble of DSMs with hybrid method	An ensemble of heterogeneous DSMs for portrait video segmentation with hybrid method was proposed and experimented. The proposed method showed more than 1.6% higher accuracy than single DSMs on average while keeping equal segmentation speed to a single DSM.

In addition to the above accomplishments, image processing and PS using downscaled images were evaluated and analysed. The downscaling of images improved processing speed 2x, while keeping the accuracy more than 90%.

This work constructed a new dataset for PVS because currently available video datasets are not dedicated for PVS. The new dataset consists of approximately 35,000 ~ 40,000 video frames with 480 × 360 resolutions showing a single person view, mainly upper body part. The dataset was collected from publicly available websites and annotated using various tools and it was verified manually.

Through this research work, three journal papers were published to reputed journals and two conference papers were presented on IEEE international conferences.

8. ANALYSIS/SIMULATION/TESTING/VALIDATION

A. Experiment setup for objective 1

The objective 1 of this work uses PortraitNet (PN), SINet (SN), FCN, MobileNetV2 (MNV2), MobileNetV3 (MNV3) and the CEDN deep-learning models for the experiment. Table 5 shows the DSM names, the dataset used for training the model, the input image resolution of the model and the dataset used for testing the model.

Table 5. Training and testing dataset of models.

DSM	Training Dataset	Resolution	Testing Dataset
CEDN	AISegment	224 × 224	EG1800 + CDI
FCN	PASCAL VOC 2011	500 × 500	EG1800 + CDI
MobileNetV2	Custom	128 × 128	EG1800 + CDI
MobileNetV3	Custom	224 × 224	EG1800 + CDI
PortraitNet	EG1800	224 × 224	EG1800 + CDI
SINet	EG1800	224 × 224	EG1800 + CDI

The AISegment dataset, which is publicly available for PS training and testing, was used to train the CEDN model. The Custom dataset consists of 18,698 portrait images which are publicly available on websites and was used by the MobileNetV2 and MobileNetV3 pre-trained models. The EG1800 dataset consists of 1447 publicly available images for training and 289 images for validation. In this paper, 270 images from the validation set of EG1800 dataset and 298 validation portrait image datasets collected from publicly available websites (called “CDI” dataset) were used for the experiment.

B. Experiment setup for objective 2 and 3

The objective 2 and 3 of this work use four DSMs for the ensemble. There are several pre-trained DSMs for the PS on open-source websites. These DSMs can be used for PS experiments immediately without any training or parameter tuning process. This paper chooses four state-of-the-art pre-trained DSMs which are publicly available. These DSMs were verified by measuring the accuracy and comparing with other papers [29]. This work uses the EG1800+CDI test dataset for the verification, where EG1800 is a dataset used in reference papers. Among the four DSMs, MNV2 uses MobileNetV2 as an encoder and an upsampling block with a transpose convolution as a decoder. Its input/output resolution is 128×128. MNV3 uses MobileNetV3 with a dept multiplier 0.5 as an encoder and an upsampling block with a transpose convolution as a decoder. It has 224×224 input/output resolution. PN is a pre-trained model of PortraitNet [28]. It has 224×224 input/output resolution. SN is a pre-trained model of SINet [29]. It has 320×320 input/output resolution. Table 6 shows the backbone of each model and the accuracy comparison using the verification dataset EG1800+CDI.

Table 6. Experimented DSMs.

Model	Backbone	IoU (%) of reference	IoU (%) of verification
MNV3	MobileNetV3	94.19	94.76
MNV2	MobileNetV2	*95.76	96.21
PN	MobileNetV2	95.99	95.7
SN	Custom	95.29	95.43

(* The value was calculated based on the accuracy ratio to SINet [29].)

This work conducted the experiment with these four pre-trained DSMs to segment the body of a human from video dataset. To the best of our knowledge, available video datasets are not dedicated to PS purpose. For the experiment, a new test video dataset was collected from publicly available websites such as News, Talk Show, Interview, etc. The dataset is a collection of captured frames from 81 videos showing a single person upper body. The resolution of videos is 480×360 pixels, the duration is 10 seconds ~ 20 seconds, and the frame rate of videos is in the range of 25 to 30 frames per second (FPS). The dataset was annotated using various tools and the annotated result was manually verified. Finally, the test dataset consists of around 40,000 frame images and ground truth masks. Single models, NM ensemble, NF ensemble, and hybrid models are evaluated using the test dataset. This work experimented with all ensembles of four single models. The simple soft voting is the combining method for the ensemble of single models. The experiment was conducted using Python with Keras, PyTorch, OpenCV and TensorFlow libraries on Ubuntu operating system and GeForce GTX 1080 GPU hardware.

C. Validation

Table 7 shows the validation of datasets used in other works and this work.

Table 7. Validation of datasets

		Related works	This work	Note
General purpose (10 papers)	**Dataset**	Small number of images (2~11 images)	800 portrait images	**- No standard dataset.**
	Key point	Information preservation or efficiency	Information preservation + speed improvement	**- Our novelty.**
Deep-Learning (11 papers since 2016)	**Dataset**	EG1800(5), Own(3), PFCN(2), Pascal VOC(2), COCO(2), OCHuman(1), Baidu(1), Supervise.ly(1)	EG1800, Own	**- EG1800** is the most popular. **- Own dataset** is also used frequently.
	Key point	Downscaled image for high-speed	Downscaled image for high-speed	**- Downscaling is a common for speed improvement.**

9. RESULTS AND DISCUSSIONS

A. Results analysis

Table 8 shows the IoU comparison of PS results for a single model without ensemble, the Two-Models ensemble using simple soft voting, the Two-Models ensemble using weighted soft voting, the Three-Models ensemble using simple soft voting and the Three-Models ensemble using weighted soft voting. It is observed that the MNV2 model, which has the highest IoU before ensemble, shows the highest IoU in Two-Models ensemble using weighted soft voting. The FCN model which has the lowest IoU before ensemble shows the highest IoU in the Two-Models ensemble using simple soft voting, as well as in the Three-Models ensemble. It shows the most improvement after ensemble, i.e., 95.2268% of IoU before ensemble is enhanced to 97.1839% after ensemble. The result shows that our approach to improve the accuracy does not degrade the accuracy of models with higher IoU, and at the same time improves the accuracy of models with lower IoU. Fig. 5 is the graphical representation of IoU comparison of individual models for no ensemble, the Two-Models ensemble and the Three-Models ensemble. It is observed that the Three-Models ensemble shows the best result compared to no ensemble and the Two-Models ensemble, while the Two-Models ensemble shows a better result than no ensemble.

Table 8. Intersection over Union (IoU) value comparison. (%)

Model	Without Ensemble	Two-Models Ensemble		Three-Models Ensemble	
		Simple Soft Voting	Weighted Soft Voting	Simple Soft Voting	Weighted Soft Voting
CEDN	95.3317	95.8407	96.1651	96.8910	96.9199
FCN	95.2268	**96.3539**	96.3768	**97.1404**	**97.1839**
MNV2	**96.2082**	96.2588	**96.4747**	97.0082	97.0369
PN	95.6992	96.0736	96.2487	96.8629	96.9233
SN	95.4326	95.8305	96.1167	96.8906	96.9394

(The best results are highlighted in **bold**.)

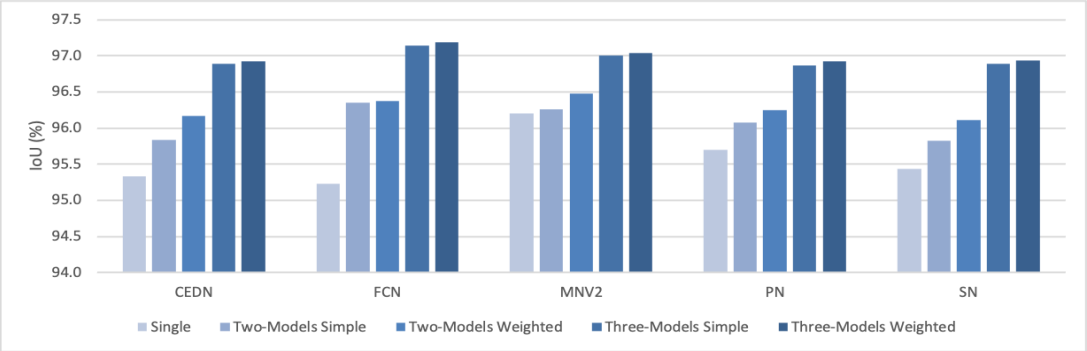


Figure 5. IoU comparison of single models and ensemble models.

Fig. 6 shows the segmented results of single models and proposed ensemble models. The first column is the original image, the second column is ground truth segmentation, followed by single models' segmentation, and the last column is the proposed ensemble model's segmentation.

Figure 6. Segmented results of single models and Two-Models ensemble: (a) original image; (b) ground truth; (c) single models; and (d) proposed ensemble model.

Table 9 shows the average values of IoU, IoU standard deviation, FNR+FDR and |FNR-FDR| from single, Two-Models (1F) NM ensemble, Two-Frames (2F) NF ensemble, Three-Models (1F) NM ensemble, and Three-Frames (3F) NF ensemble models. Two-Models (1F) NM ensemble and Two-Frames (2F) NF ensemble models significantly improve IoU, IoU std, FNR+FDR, and |FNR-FDR| than single models. The difference of IoU value between Two-Models (1F) NM ensemble and Two-Frames (2F) NF ensemble models is 0.0615%. Three-Models (1F) NM ensemble and Three-Frames (3F) NF ensemble also show significant improvement in IoU, IoU std, FNR+FDR, and |FNR-FDR| than single models. The difference of IoU value between Three-Models (1F) NM ensemble and Three-Frames (3F) NF ensemble models is 0.1933%.

Table 9. Average of all models (%)

	IoU	IoU Std	FNR+FDR	|FNR-FDR|
Single Models	92.9653	2.1149	7.2717	1.6290
Two-Models (1F)	95.1868	1.8978	4.9354	1.4894
Two-Frames (2F)	95.1253	1.9385	5.0001	1.4980
Three-Models (1F)	95.3667	1.6814	4.7386	0.8465
Three-Frames (3F)	95.1734	1.8182	4.9428	0.8473

Fig. 7 compares IoU and FLOPs of single models, NM ensemble, and NF ensemble models. There are four groups in the chart. Group#1 is the upper right part of the chart, and it shows high accuracy and high use of computing power. MNV3, MNV3+MNV2+SN (1F), and MNV2+SN (1F) models belong to this group. Group#2 is the lower right part of the chart, and it shows high accuracy and low use of computing power. SN, MNV2+SN (2F), MNV3+MNV2+SN (3F), MNV2+PN+SN (3F), and MNV3+SN (2F) models belong to this group. Group#3 is the lower left part of the chart, and it shows a low accuracy and low use of computing power. PN and MNV3 models belong to this group. In general, a model that produces high accuracy using low computing power is desirable, and the models of Group#2 belong to this category. The

result shows that proposed NF ensemble models can perform high accuracy using low computing power than other single and NM ensemble models.

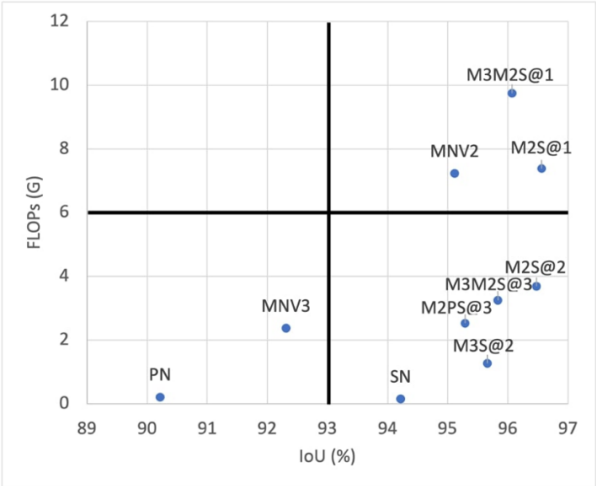


Figure 7. Comparison of IoU and FLOPs; M3M2S@1 = MNV3+MNV2+SN (1F); M2S@1 = MNV2+SN (1F); M2S@2 = MNV2+SN (2F); M3M2S@3 = MNV3+MNV2+SN (3F); M2PS@3 = MNV2+PN+SN (3F); M3S@2 = MNV3+SN(2F).

Table 10 and Fig. 8 show the average of MER and CER of single models, NM ensemble, and NF ensemble models. All ensemble models show higher IoU values than a single model. Two-Frames (2F) NF ensemble and Three-Frames (3F) NF ensemble show lower CER than a single model. It indicates that proposed NF ensemble models can perform better accuracy in a cost-efficient way than single models. However, it is also observed that the requirement of memory for the ensemble is increased as the number of models for the ensemble is increased.

Table 10. Average of MER and CER

Model	IoU (%)	MER	CER
Single model	92.965	1.885	2.640
Two-Models (1F)	95.187	3.685	5.211
Two-Frames (2F)	95.125	3.688	2.607
Three-Models (1F)	95.367	5.517	7.817
Three-Frames (3F)	95.173	5.528	2.611

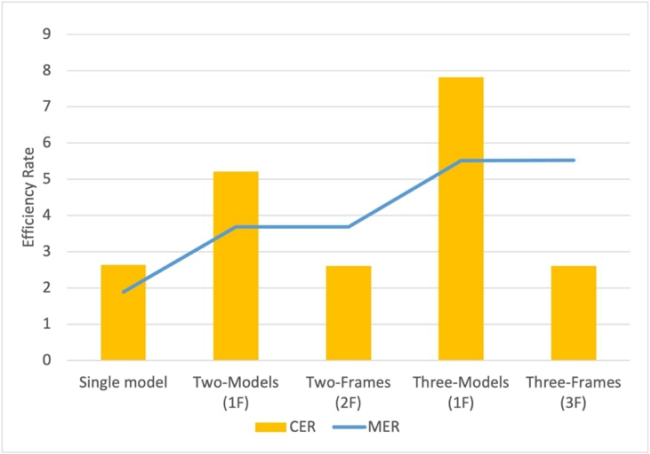


Figure 8. Average of CER and MER

Fig. 9 shows the PS result of single models and the proposed NF ensemble model. Column (a) is the original frames of a video; (b) and (c) are the segmentation results using single models; (d) is the result using the proposed NF ensemble; and (e) is ground truth. It shows the segmentation result of continuous frames. The proposed NF ensemble shows stable segmentation result while single models show irregular segmentation results between adjacent frames.

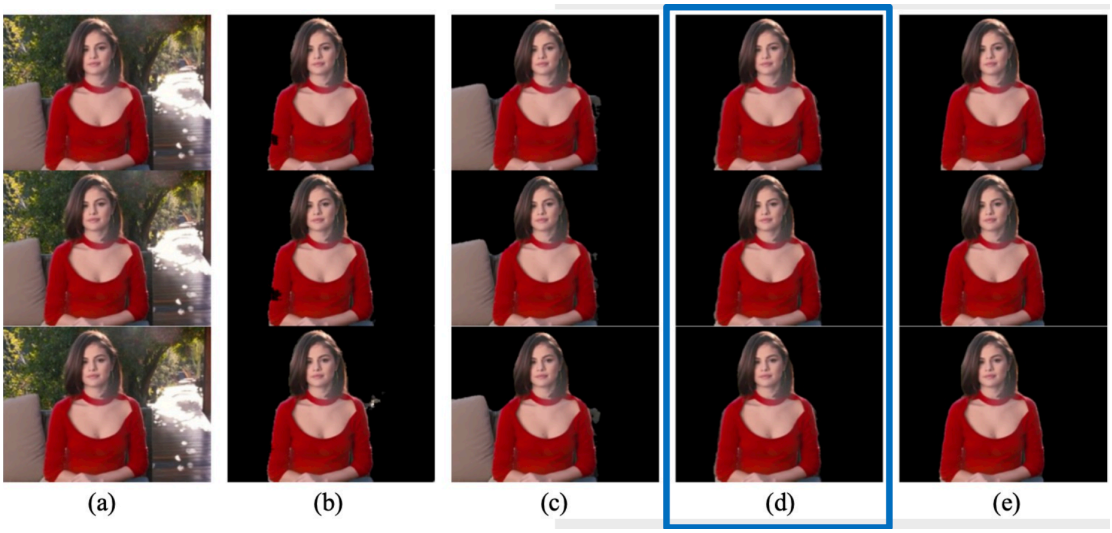


Figure 9. (a) Input frames; (b) output of Segmentation mode 1; (c) output of Segmentation model 2; (d) output of proposed N-Frames Ensemble; (e) expected output (ground truth)

Table 11 shows the IoU values of Parallel-BMDS approach. When a DSM takes 5 frame times for inference, the proposed approach can produce the segmentation results without missing any video frames during the inference time. At the same time, it can perform with 99% accuracy rate to the DSM.

Table 11. IoU values of Parallel-BMDS

		DSM frame time			
		2	3	4	5
BMDS frames	w/o	94.727	94.727	94.727	94.727
	0	94.685	94.368	94.063	93.758 (99%)
	1	94.577	94.267	93.946	93.661
	2	94.471	94.147	93.843	93.554
	3	94.348	94.040	93.734	93.473
	4	94.237	93.921	93.657	93.377
	5	94.116	93.844	93.550	93.297

Table 12 shows the IoU values of Ensemble-BMDS approach. The propose approach produces higher accuracy than single DSM, at the same time it can process all video frames to generate segmentation results. Fig. 10 show the speed comparison of a single DSM and BMDS method.

Table 12. IoU values of Ensemble-BMDS

		DSM frame time			
		2	3	4	5
BMDS frames	w/o	94.727	94.727	94.727	94.727
	0	**96.335**	**95.990**	**95.644**	**95.309**
	1	**96.215**	**95.863**	**95.515**	**95.191**

	2	**96.086**	**95.733**	**95.390**	**95.078**
	3	**95.950**	**95.605**	**95.276**	**94.967**
	4	**95.821**	**95.480**	**95.169**	**94.864**
	5	**95.696**	**95.375**	**95.055**	**94.755**

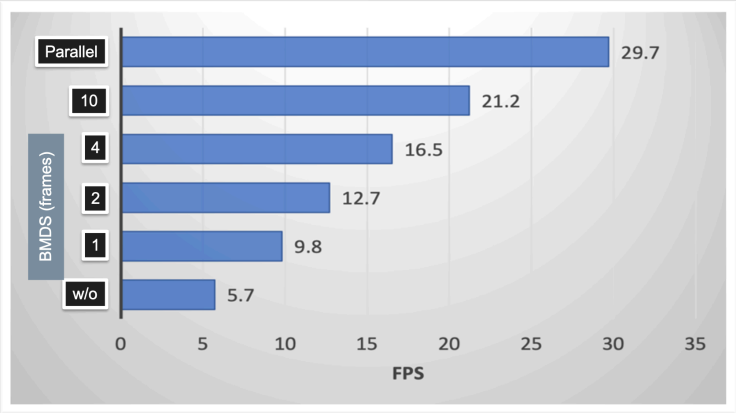
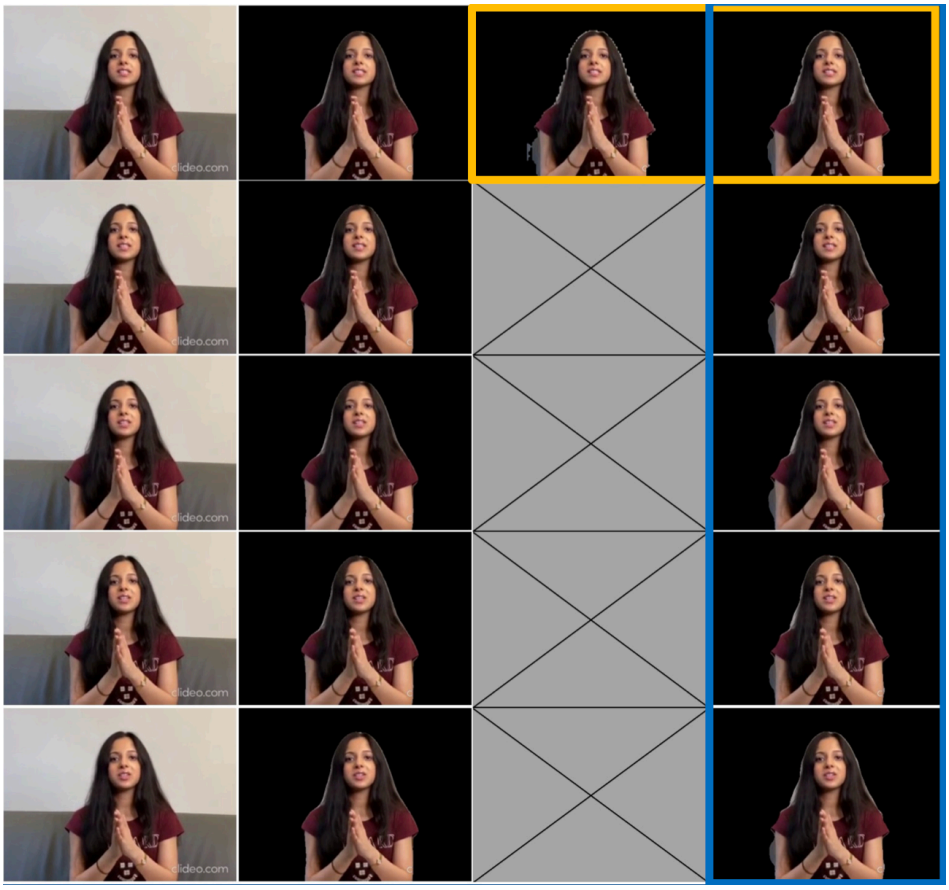


Figure 10. Speed comparison

Fig. 11 shows a segmentation example of continuous video frames using the Parallel-BMDS. The result shows that the proposed hybrid method produces segmentation result without missing any video frames when the inference of a DSM takes longer time than a single frame time.



(a) (b) (c) (d)

Figure 11. Segmentation examples of continuous video frames (5 frames); (a) video frames; (b) ground truth; (c) segmentation using a slow DSM; (d) segmentation using Parallel-BMDS

[Additional experiment]

This work performed an extensive experiment to find out a relationship between the layer/layer structure of a DSM model and the result of ensemble model. It was fulfilled by measuring Stacking ensemble results with/without the layer modification of a DSM. For the experiment, the network layer of the CEDN model was modified in various way and the models were trained with a same dataset (EG18000) and same training parameters (learning rate = 0.0001, epochs = 30). According to the extensive experiment result, the followings are observed:

- CEDN models having similar IoU value show similar improvement after the ensemble, no matter which layers of the CEDN model were modified.
- All ensemble models show similar pattern, no matter which layers were modified, and no matter which DSMs were combined for the ensemble.
- There is no significant difference of the ensemble result between the modified CEDN models and original CEDN models trained using different parameters (30 epochs and 50 epochs).
- CEDN models trained using different datasets show similar pattern to the ensemble results of the modified CEDN models, no matter which datasets were used, and no matter which DSMs were combined for the ensemble.
- The accuracy of ensemble models has dependency only with the accuracy of models involved in the ensemble.
- When a CEDN model has lower IoU value, it shows higher improvement after the ensemble with other DSMs.
- It is hard to find any relationship between network layer difference and the accuracy of experimented ensemble models.

According to the results, we found that it was hard to find any relationship between network layer and the accuracy of ensemble model. The results show that the accuracy of Stacking ensemble models has dependency only with the accuracy of models involved in the ensemble.

The execution time of PS has many dependencies to a hardware, compiler, optimization parameters, etc. The execution times of four DSMs (MNV2, MNV3, PN, SN) from the original papers are approximately 24 millisecond (ms), 63ms, 32ms, and 10ms each. This work measured execution time of these models, and it showed approximately 83.5ms, 83.5ms, 35ms, and 10ms each on the mobile device. Table 13 shows the comparison of execution time of single DSMs and propose approaches.

Table 13. Comparison of execution time (millisecond)

	MNV2	MNV3	PN	SN	NF-Ensemble	Hybrid (5f)
Reference	24	63	32	10	X	X
Ours	83.5	83.5	35	10	53	10

B. Discussions

For the first objective, this work proposed an ensemble of heterogeneous DSMs for portrait image segmentation, and it achieved more than 1.4% higher accuracy than single DSMs on average. For the second

objective, a novel N-Frames ensemble of multiple DSMs for PVS was introduced, and it performed with more than 2% higher accuracy than single DSMs on average. It also showed lower variance and bias errors than single DSMs. The average computing power of the proposed approach is 2.49 Giga FLOPs, and it is same with single DSMs. The result shows that the proposed method is as fast as a single DSM, at the same time, it can produce higher accuracy than a single DSM. For the third objective, a novel PS method using the ensemble of DSMs and hybrid method for PVS was proposed. The proposed approach produced 1.6% higher accuracy than single DSMs on average while keeping same segmentation speed to a single DSM, and it achieved 5x faster segmentation speed than a single DSM while keeping approximately 99% accuracy rate compared to a single DSM. It also performed with approximately 30 Frames per seconds (FPS) speed of PVS on Nvidia Jetson-Nano hardware, whereas the segmentation speed of single DSM was 5.7 FPS.

The experiment results show that the proposed ensemble approaches improve accuracy, variance and bias errors of PS and PVS compared to single DSMs involved in the ensemble. The hybrid approach using the BMDS method shows higher segmentation speed than single DSMs, at the same time, it produces reasonable segmentation results.

10. CONCLUSIONS

The applications using PS technology are increasing more and more. Several studies on PS using deep-learning models and object segmentation using the ensemble of DSMs were reported from 2016. However, the PS using ensemble approach is an area that needs further study. This paper introduced three ensemble approaches of multiple DSMs for high-performance PS. The advantages of the proposed ensemble approaches are 1) it can generate optimized results using an ensemble of multiple DSMs. As a result, the accuracy of segmentation result is higher than sing DSMs involved in the ensemble. 2) The first approach can produce higher accuracy, lower variance, and lower bias error than single DSMs. 3) The second approach performs with higher accuracy than single DSMs, at the same time it is as fast as a single DSM. It is also robust to the sudden noises of a video. 4) The third approach performs with significantly faster segmentation speed than a single DSM, at the same time it produces higher accuracy than single DSMs. The experiment result shows that the proposed first ensemble approach performs with more than 1.4% higher accuracy than sing DSMs for portrait image segmentation. The second ensemble approach produces more than 2% higher accuracy than single DSMs for PVS, at the same time, it consumes equal computing power to single DSMs. The third ensemble approach achieves more than 1.6% higher accuracy than single DSMs while keeping same segmentation speed, and it also performs with 5x faster segmentation speed than a single DSM while keeping approximately 99% accuracy rate compared to a single DSM. This work can be utilized for a selfie photo segmentation in mobile camera apps, a virtual background function of video conferencing solutions, a robot vision domain for human-robot interaction and contents-based image/video retrieval in database systems, etc. Future work is to apply the proposed approaches to the segmentation of other objects other than human body.

BIBLIOGRAPHY

[1] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, “A review of semantic segmentation using deep neural networks,” Int. J. Multimed. Inf. Retr., vol. 7, no. 2, pp. 87–93, 2018, doi: 10.1007/s13735-017-0141-z.

[2] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image segmentation using Deep Learning: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PP, pp. 1–1, 2021.

[3] R. M. Haralick and L. G. Shapiro, “Image segmentation techniques.,” Comput. Vision, Graph. Image Process., vol. 29, no. 1, pp. 100–132, 1985, doi: 10.1016/S0734-189X(85)90153-7.

[4] I. Ahmed, M. Ahmad, F. A. Khan and M. Asif, "Comparison of Deep-Learning-Based Segmentation Models: Using Top View Person Images," in *IEEE Access*, vol. 8, pp. 136361-136373, 2020, doi: 10.1109/ACCESS.2020.3011406.

[5] J. Yang, B. Price, S. Cohen, H. Lee, and M. H. Yang, “Object contour detection with a fully convolutional encoder-decoder network,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 193–202, 2016, doi: 10.1109/CVPR.2016.28.

[6] X. Liu, Z. Deng, and Y. Yang, “Recent progress in semantic image segmentation,” Artif. Intell. Rev., vol. 52, no. 2, pp. 1089–1106, 2019.

[7] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, “A survey on deep learning techniques for image and video semantic segmentation,” Appl. Soft Comput., vol. 70, pp. 41–65, 2018.

[8] M. Gruosso, N. Capece, and U. Erra, “Human segmentation in surveillance video with deep learning,” Multimed. Tools Appl., vol. 80, no. 1, pp. 1175–1199, 2021.

[9] J. Zhang, X. Zhao, Z. Chen and Z. Lu, "A Review of Deep Learning-Based Semantic Segmentation for Point Cloud," in IEEE Access, vol. 7, pp. 179118-179133, 2019, doi: 10.1109/ACCESS.2019.2958671.

[10] B. Zhao, J. Feng, X. Wu, and S. Yan, “A survey on deep learning-based fine-grained object classification and semantic segmentation,” Int. J. Autom. Comput., vol. 14, no. 2, pp. 119–135, 2017, doi: 10.1007/s11633-017-1053-3.

[11] L. Rokach, “Ensemble-based classifiers,” Artif. Intell. Rev., vol. 33, no. 1–2, pp. 1–39, 2010, doi: 10.1007/s10462-009-9124-7.

[12] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. Cambridge (EE. UU.): MIT Press, 2016.

[13] L. Breiman, “Stacked regressions,” Mach. Learn., vol. 24, no. 1, pp. 49–64, 1996, doi: 10.1023/A:1018046112532.

[14] L. K. Hansen and P. Salamon, “Neural Network Ensembles,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. 10, pp. 993–1001, 1990, doi: 10.1109/34.58871.

[15] Y. Zuo and T. Drummond, “Fast Residual Forests: Rapid Ensemble Learning for Semantic Segmentation,” Proc. 1st Annu. Conf. Robot Learn. PMLR 78, no. CoRL, pp. 27–36, 2017, [Online]. Available: http://proceedings.mlr.press/v78/zuo17a.html.

[16] Y. Koren, “The BellKor Solution to the Netflix Grand Prize,” no. August, pp. 1–10, 2009.

[17] J. Luiten, P. H. S. Torr, and B. Leibe, “Video Instance Segmentation 2019 : A winning approach for combined Detection , Segmentation , Classification and Tracking .,” 2019 IEEE/CVF Int. Conf. Comput. Vis. Work., pp. 709–712, 2019, doi: 10.1109/ICCVW.2019.00088.

[18] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia, “Deep automatic portrait matting,” in Computer Vision – ECCV 2016, Cham: Springer International Publishing, 2016, pp. 92–107.

[19] X. Du et al., “Boundary-sensitive network for portrait segmentation,” Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2019, pp. 1–8, 2019, doi: 10.1109/FG.2019.8756516.

[20] B. Lyu, Y. Yang, S. Wen, T. Huang, and K. Li, “Neural architecture search for portrait parsing,” IEEE Trans. Neural Netw. Learn. Syst., vol. PP, 2021.

[21] J. Miao, K. Sun, X. Liao, L. Leng, and J. Chu, “Human segmentation based on compressed deep convolutional neural network,” IEEE Access, vol. 8, pp. 167585–167595, 2020.

[22] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, “Boundary-aware feature propagation for scene segmentation,” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[23] X. Shen et al., “Automatic portrait segmentation for image stylization,” Comput. Graph. Forum, vol. 35, no. 2, pp. 93–102, 2016, doi: 10.1111/cgf.12814.

[24] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Computer Vision – ECCV 2018, Cham: Springer International Publishing, 2018, pp. 833–851.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[27] A. Howard et al., “Searching for MobileNetV3,” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[28] S.-H. Zhang, X. Dong, H. Li, R. Li, and Y.-L. Yang, “PortraitNet: Real-time portrait segmentation network for mobile device,” Comput. Graph., vol. 80, pp. 104–113, 2019.

[29] H. Park, L. L. Sjosund, Y. Yoo, N. Monet, J. Bang, and N. Kwak, “SINet: Extreme lightweight portrait segmentation networks with spatial squeeze modules and information blocking decoder,” in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020.

[30] X. Shen et al., “Automatic portrait segmentation for image stylization,” Comput. Graph. Forum, vol. 35, no. 2, pp. 93–102, 2016.

[31] S. K. Warfield, K. H. Zou, and W. M. Wells, “Simultaneous Truth and Performance Level Estimation (STAPLE): An algorithm for the Vaidation of Image Segmentation,” IEEE Trans. Med. Imaging, vol. 23, no. 7, pp. 903–921, 2004, doi: 10.1109/TMI.2004.828354.

[32] A. Holliday, M. Barekatain, J. Laurmaa, C. Kandaswamy, and H. Prendinger, “Speedup of deep learning ensembles for semantic segmentation using a model compression technique,” Comput. Vis. Image Underst., vol. 164, pp. 16–26, 2017, doi: 10.1016/j.cviu.2017.05.004.

[33] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, “Semantic Segmentation of Aerial Images With an Ensemble of Cnns,” ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci., vol. III–3, no. July, pp. 473–480, 2016, doi: 10.5194/isprsannals-iii-3-473-2016.

[34] J. Li, S. He, H. Wong and S. Lo, "Proposal-Driven Segmentation for Videos," in IEEE Signal Processing Letters, vol. 26, no. 8, pp. 1098-1102, Aug. 2019, doi: 10.1109/LSP.2019.2921654.

[35] Y. Liu, C. Shen, C. Yu, and J. Wang, “Efficient semantic video segmentation with per-frame inference,” in Computer Vision – ECCV 2020, Cham: Springer International Publishing, 2020, pp. 352–368.

[36] S. Lin, F. Qin, H. Peng, R. A. Bly, K. S. Moe, and B. Hannaford, “Multi-frame feature aggregation for real-time instrument segmentation in endoscopic video,” IEEE Robot. Autom. Lett., vol. 6, no. 4, pp. 6773–6780, 2021.

[37] Y. Wang, W. Zhang, L. Wang, F. Yang, and H. Lu, “Temporal consistent portrait video segmentation,” Pattern Recognit., vol. 120, no. 108143, p. 108143, 2021.

[38] T. Zhang, C. Lang, and J. Xing, “Realtime Human Segmentation in Video,” in MultiMedia Modeling, Cham: Springer International Publishing, 2019, pp. 206–217.

PUBLICATIONS

I. Journals

1. Y.-W. Kim, Y.-C. Byun, A. V. N. Krishna and B. Krishnan, "Selfie Segmentation in Video Using N-Frames Ensemble," in IEEE Access, vol. 9, pp. 163348-163362, 2021, doi: 10.1109/ACCESS.2021.3133276. (SCIE & WoS & Scopus).
2. Y.-W. Kim, Y.-C. Byun, and A. V. N. Krishna, “Portrait segmentation using ensemble of heterogeneous deep-learning models,” Entropy (Basel), vol. 23, no. 2, p. 197, 2021, doi: 10.3390/e23020197. (SCIE & WoS & Scopus).
3. Y. W. Kim, I. Rose, and A. V. N. Krishna, “A study on the effect of canny edge detection on downscaled images,” Pattern Recognit. Image Anal., vol. 30, no. 3, pp. 372–381, 2020. (Scopus).

II. Conferences

1. Y. W. Kim, J. Innila Rose and A. V. N. Krishna, "Accuracy Enhancement of Portrait Segmentation by Ensembling Deep Learning Models," 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2020, pp. 59-64, doi: 10.1109/ICRCICN50933.2020.9296196. (IEEE Xplore & Scopus).
2. Y. W. Kim, A. R. Oh, I. Rose J and A. V. N. Krishna, "Analyzing the Performance of Canny Edge Detection on Interpolated Images," 2019 International Conference on Information and Communication Technology Convergence (ICTC), 2019, pp. 726-730, doi: 10.1109/ICTC46691.2019.8939595. (IEEE Xplore & WoS & Scopus)